Particle Filter Inference in an Articulatory-Based Speech Model

Beierholm, Thomas; Winther, Ole

Published in:
I E E E Signal Processing Letters

Link to article, DOI:
10.1109/LSP.2007.899332

Publication date:
2007

Document Version
Early version, also known as pre-print

Particle Filter Inference in an Articulatory Based Speech Model

Thomas Beierholm (*), *Student Member, IEEE,* and Ole Winther

***Abstract*— A time-varying auto-regressive speech model parameterized by formant frequencies, formant bandwidths and formant gains is proposed. Inference in the model is made by particle filtering for the application of speech enhancement. The advantage of the proposed parametrization over existing parameterizations based on AR coefficients or reflection coefficients is the smooth time-varying behavior of the parameters and their loose coupling. Experiments confirm this advantage both in terms of parameter estimation and SNR improvement. Finally, further modelling and inference improvements are outlined.**

***Index Terms*— Particle filtering, time-varying auto-regressive speech model, formant frequency.**

I. Introduction

In the application of speech enhancement, the speech signal is commonly modelled as a time-varying Auto-Regressive (AR) Gaussian process. In block-processing systems the speech signal is assumed quasi-stationary meaning that the parameters of the AR process describing the speech signal are assumed fixed in the duration of the block. As described in Ref. [1] the articulators of speech, such as the vocal tract, are continually moving, hence the assumption of quasi-stationarity of speech can be improved upon. The Time-Varying Auto-Regressive (TVAR) model used in Refs. [1], [2] lets the parameters of the AR process describing the speech signal vary from sample to sample and thus avoids the assumption of quasi-stationarity of the speech signal.

The TVAR model facilitates a state-space formulation of the observed noisy signal in which the problem of joint estimation of the unknown parameters of the model and the state sequence becomes a challenge. One approach is to perform ML estimation using the EM algorithm. A different approach was used in Refs. [1], [2], where sequential Bayesian estimation of the unknown parameters and state sequence was performed by particle filtering. Instead of using the AR coefficients directly, then, in Ref. [2] the TVAR model was reparameterized in terms of reflections coefficients as this lead to a stronger physical interpretation of the model and stability of the model could easily be verified.

In this paper a reparametrization of the TVAR model with an even stronger physical interpretation is used [3]. The TVAR model is parameterized in terms of formant frequencies, formant bandwidths and formant gains, called the fbg parameters in the following. It is intended that this new parametrization can lead to improved particle filtering by way of exploiting known properties of the fbg parameters and thereby eventually improve quality of the estimated speech signal. As stressed in Ref. [3] the new parameters have a slow time variation due to the inertia of the speech producing system in contrast to the reflection coefficients which can have a rapid time variation. The new parameters are also loosely coupled and exhibit smooth trajectories and stability of the model is easily ensured.

A common feature of the TVAR model [1] and the fbg parameterized model introduced in this work is that conditional on the unknown parameters of the model, the model reduces to a linear Gaussian state-space system. In Ref. [1] this feature was made use of in a variance reduction (Rao-Blackwellization) step whereby the problem of sampling from the joint posterior distribution of the states and the unknown parameters of the model is reduced to that of sampling from the posterior distribution of the unknown parameters only.

The main idea behind the work described in this paper is to introduce a TVAR model with an even stronger physical interpretation than the reflection coefficients [2] and obtain filtered estimates of the clean speech signal with Rao-Blackwellized particle filtering [1].

II. Time-Varying Auto-Regressive Model

In the Time-Varying Auto-Regressive (TVAR) model a speech signal is modelled as a non-stationary AR(p) process, where p denotes the order of the AR process which is assumed fixed in the following. The coefficients of the AR(p) process and the variance of the process noise are allowed to change from sample to sample, i.e.

$$x[n] = \sum_{i=1}^{p} a_i[n]x[n-i] + \sigma_e[n]e[n] \quad , \quad e[n] \sim \mathcal{N}(0,1) \ ,$$

where $e[n]$ is the innovation sequence with variance $\sigma_e^2[n]$ and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is assumed that the speech signal is contaminated by non-stationary Gaussian noise

$$y[n] = x[n] + \sigma_d[n]d[n] \quad , \quad d[n] \sim \mathcal{N}(0,1) \ ,$$

where $d[n]$ denotes the observation noise with variance $\sigma_d^2[n]$. The TVAR model is conveniently formulated as a state-space model with the following state and observation equations

$$\mathbf{x}[n] = \mathbf{A}[n]\mathbf{x}[n-1] + \mathbf{B}[n]\mathbf{v}[n] \ , \ \mathbf{v}[n] \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{I}_p) \quad (1)$$

$$y[n] = \mathbf{C}\mathbf{x}[n] + \mathbf{D}[n]\mathbf{w}[n] \ , \ \mathbf{w}[n] \sim \mathcal{N}(0,1) \ , \quad (2)$$

where $\mathbf{a}[n] = (\, a_1[n], \, \cdots \, , \, a_p[n] \,)^T$ is the coefficient vector, $\mathbf{x}[n] = (\, x[n], \, \cdots \, , \, x[n-p+1] \,)^T$ the state vector and

$$\mathbf{A}[n] = \begin{pmatrix} \mathbf{a}^T[n] \\ \mathbf{I}_{(p-1)} \quad \mathbf{0}_{(p-1)\times 1} \end{pmatrix} \ , \ \mathbf{B}[n] = \begin{pmatrix} \sigma_e[n] \\ \mathbf{0}_{(p-1)\times 1} \end{pmatrix} \quad (3)$$

Manuscript received July xx, 2006; revised Month xx, 2006.

T. Beierholm is with GN ReSound A/S, Lautrupbjerg 7, P.O. Box 99, 2750 Ballerup, Denmark, email tbe@imm.dtu.dk.

O. Winther is with IMM, Danish Technical University, 2800 Lyngby, Denmark.

$$\mathbf{C} = \begin{pmatrix} 1\,0\cdots 0 \end{pmatrix} \ , \ \mathbf{D}[n] = (\ \sigma_d[n]\ ) \ . \tag{4}$$

The state-space formulation of the TVAR model in eqs. (1) and (2), with the parametrization eqs. (3) and (4), is used in Ref. [1]. The unknown parameters of the model are the p AR coefficients in $\mathbf{a}[n]$ and the innovation and observation noise variances. The AR coefficients and the two noise variance parameters represented by their logarithms were assumed independent and taken as evolving according to first-order Markov random walk processes. The variance of the random walk processes for each of the AR coefficients and the variances of the random walk processes for the logarithms to the variance of the innovation sequence and observation noise are denoted δ_a, δ_e, δ_d, respectively.

III. Articulatory-Based Speech Model

The formulation of a speech model based on the fbg parameters, which are close to the articulators of speech [3], is based on what in the area of speech synthesis, is referred to as a *Parallel Formant Synthesizer* (PFS) and therefore the model is called the PFS model in the following. A PFS synthesizes speech by summing the outputs of a number of parallel connected resonance circuits. The structure of a PFS is illustrated in Fig. 1. The resonators are driven by a common excitation signal which is taken to be white standard normal distributed noise. Each resonance circuit models a formant in the spectrum of the speech signal, in the sense that the spectrum of the excitation signal is shaped to have a peak at the resonance frequency and the bandwidth and gain of the 'bump' is determined by the resonance circuit as well. The resonators are taken as second-order IIR filters with z-transforms

$$H_k(z) = \frac{g_k}{1 + a_{k,1}(f_k, b_k)z^{-1} + a_{k,2}(b_k)z^{-2}} \ , \tag{5}$$

where f_k, b_k and g_k denotes the formant frequency, formant bandwidth and formant gain, respectively, of the k^{th} formant. The mapping from these parameters to the coefficients of the resonators is given by [4], [5]

$$a_{k,1}(f_k, b_k) = 2\exp(-\pi b_k/f_s)\cos(2\pi f_k/f_s) \tag{6}$$

$$a_{k,2}(b_k) = -\exp(-2\pi b_k/f_s) \ , \tag{7}$$

where f_s is the sampling frequency in Hz. By letting

$$\mathbf{A}_k[n] = \begin{pmatrix} a_{k,1}(f_k, b_k) & a_{k,2}(b_k) \\ 1 & 0 \end{pmatrix} \tag{8}$$

then in state-space form, the PFS model is described by the TVAR model in eqs. (1) and (2) with parametrization

$$\mathbf{A}[n] = \mathrm{diag}\,(\mathbf{A}_1, \cdots, \mathbf{A}_K) \tag{9}$$

$$\mathbf{B}[n] = \begin{pmatrix} \mathbf{0}_{1\times 2K} \\ g_1[n]\,0 \quad \cdots \quad g_K[n]\,0 \end{pmatrix}^T \tag{10}$$

$$\mathbf{C} = \begin{pmatrix} 1\,0\,1\,0\cdots 1\,0 \end{pmatrix} \ , \ \mathbf{D}[n] = (\ \sigma_d[n]\ ) \ , \tag{11}$$

where $\mathbf{x}[n] = (x_1[n],\ x_1[n-1],\ \cdots,\ x_K[n],\ x_K[n-1])^T$ is the state vector, K is the number of formants and $\mathbf{v}[n] \sim \mathcal{N}(\mathbf{0}_{2\times 1}, \mathbf{I}_2)$. A formulation based on a cascade structure of second-order sections can also be used. A description of the pros and cons of the cascade and parallel structures in speech synthesis can be found in Ref. [5]. The PFS model has $3K+1$ parameters. The f_k and b_k parameters and the logarithm to the

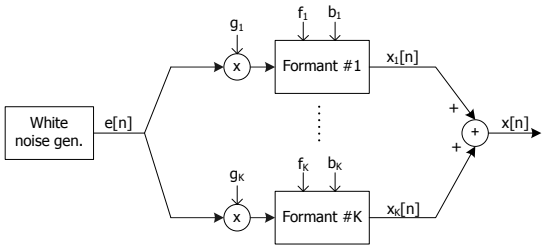

Fig. 1. Block diagram of Parallel Formant Synthesizer.

g_k parameters are assumed independent and taken as evolving according to first-order Markov random walk processes with variances δ_f, δ_b and δ_g, respectively.

IV. Particle Filter Inference

We provide a brief summary of the particle filter method used for inference in the PFS model as it is described in detail elsewhere [1]. Filtering refers to the task of computing the *filtering distribution* $p(\mathbf{a}_n, \boldsymbol{\theta}_n \,|\, \mathbf{y}_{1:n})$ and the MMSE estimate of the state vector and the parameter vector $\boldsymbol{\theta}_n$ at time n using observations $\mathbf{y}_{1:n} = (y[1],\ \cdots,\ y[n])$ up to and including time n. The joint distribution is decomposed using Bayes rule:

$$p(\mathbf{a}_n, \boldsymbol{\theta}_{0:n} \,|\, \mathbf{y}_{1:n}) = p(\mathbf{a}_n \,|\, \boldsymbol{\theta}_{0:n}, \mathbf{y}_{1:n})\, p(\boldsymbol{\theta}_{0:n} \,|\, \mathbf{y}_{1:n}) \ . \tag{12}$$

For both models (TVAR and PFS) the state vector can be integrated out analytically because $p(\mathbf{a}_n \,|\, \boldsymbol{\theta}_{0:n}, \mathbf{y}_{1:n})$ is Gaussian. This so-called Rao-Blackwellization has the effect of reducing the variance of the MMSE estimate of the state and parameter vectors. The problem is then reduced to sampling from the lower dimensional distribution $p(\boldsymbol{\theta}_{0:n} \,|\, \mathbf{y}_{1:n})$ instead of sampling from $p(\mathbf{a}_{0:n}, \boldsymbol{\theta}_{0:n} \,|\, \mathbf{y}_{1:n})$. In particle filtering this distribution is approximated by a weighted sum of δ-functions (the particles). The importance weight of particle with state space history $\boldsymbol{\theta}_{0:n}$ is given by

$$w(\boldsymbol{\theta}_{0:n}) \propto \frac{p(\boldsymbol{\theta}_{0:n} \,|\, y_{1:n})}{\pi(\boldsymbol{\theta}_{0:n} \,|\, y_{1:n})} \ , \tag{13}$$

where $\pi(.)$ denotes the importance distribution where samples are drawn from. Sequential importance sampling can be performed if the importance distribution is restricted to be of the general form

$$\pi(\boldsymbol{\theta}_{0:n} \,|\, y_{1:n}) = \pi(\boldsymbol{\theta}_{0:n-1} \,|\, y_{1:n-1})\pi(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{0:n-1}, y_{1:n}) \tag{14}$$

which facilitates recursive propagation of the importance weights in time. The crucial restriction is that the time dependence only goes to $n-1$ in the first term. Inserting eq. (14) in eq. (13) and expanding the numerator using Bayes' rule and using the assumption that the parameters evolve according to a first-order Markov process, i.e. $p(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{0:n-1})$=$p(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})$, then the weights obey $w(\boldsymbol{\theta}_{0:n})$=$w(\boldsymbol{\theta}_{0:n-1})w_n$ with

$$w_n \propto \frac{p(y_n|\boldsymbol{\theta}_{0:n}, y_{1:n-1})p(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})}{\pi(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{0:n-1}, y_{1:n})} \ . \tag{15}$$

This way sequential importance sampling avoids the need for storing the paths $\boldsymbol{\theta}_{0:n-1}$ of the particles. The complexity of recursively computing the weights can be simplified if the importance distribution at time n is set equal to the prior distribution, i.e. $\pi(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}_{0:n-1}, y_{1:n}) = p(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})$ so that (15) reduces to $w_n \propto p(y_n|\boldsymbol{\theta}_{0:n}, y_{1:n-1})$. However, this step contributes to a degeneracy whereby all weights except one after a few iterations are very close to zero. This happens because the importance distribution is different from the true posterior distribution. As a remedy, a selection step is introduced. The selection step duplicates particles in proportion to their importance weights in such a way that all particles have approximately the same weight after the selection step.

V. Experiments

The performance of the PFS model is examined and compared to the TVAR model examined by Vermaak et al. in [1]. The starting point is the TIMIT speech sentences "In simpler terms, it amounts to pointing the platform in the proper direction." (si1466) and "His sudden departure shocked the cast." (sx111) both downsampled to 16 kHz. From each sentence a 0.38s (6000 samples) speech sound is extracted. The waveforms of the extracted sounds are seen in Fig. 2. Wide-band spectrograms of the two extracted sounds were made with formant frequency tracks overlayed. From these spectrograms the following observations were made. The speech clip shown in the upper plot in Fig. 2 is voiced with a distinct pitch contour and the formants are clearly marked and changes smoothly over time. The speech clip shown in the lower plot of Fig. 2 contains both voiced and unvoiced sounds, the formants are less distinct and the formants changes less smoothly. From the wide-band spectrogram of the waveform

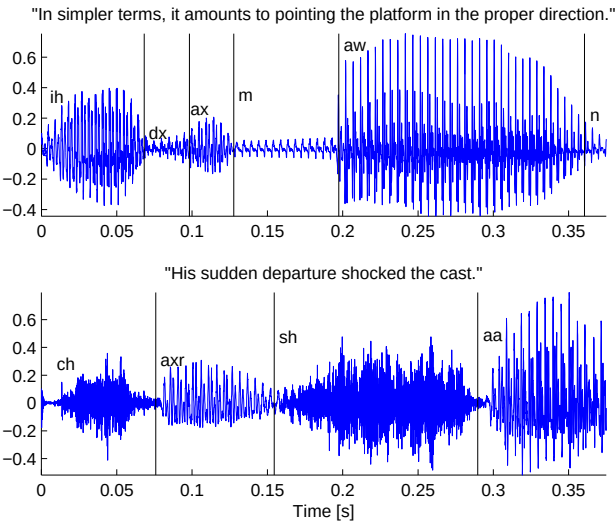

Fig. 2. Plot of the waveforms for extracted sounds from two TIMIT sentences. Phonetic information is also shown in the plots.

shown in the upper plot of Fig. 2 it was evident that by low-pass filtering $F1$ could be separated from the other formants. A modified speech sound was created where frequencies above 1100 Hz were suppressed and only $F1$ existed in the modified speech sound. In the same manner a speech sound with only $F1$ and $F2$ was created. Those two modified sounds were contaminated by zero-mean stationary white Gaussian noise at 0 dB SNR and subsequently used to manually tune the random walk parameters for both the PFS and the TVAR model so that the particle filtering gave as high SNR improvements as possible. The particle filtering used by Ref. [1] was modified in two respects; 1) so that it exploited that the variance of the observation noise was known and constant and 2) it was initialized using f and b parameters that were then mapped to AR coefficients using eqs. (6) and (7). In this way the initializations in the TVAR model and the PFS model were alike. The first 800 samples were not used in the computation of the SNR improvements in order to minimize initialization effects. The manual tuning of the random walk parameters lead to the following setting δ_a=2.5×10^{-5} and δ_e=5×10^{-3} in the TVAR model and the setting δ_f=20, δ_b=7 and δ_g=5×10^{-3} in the PFS model. Both the performance of the TVAR model and the PFS model was found to be relative insensitive to the setting of the random walk parameters.

With these settings and using 100 particles the particle filter was run on the sound where only $F1$ exists. The experiment was repeated 7 times. The TVAR model was specified to use 2 AR coefficients and in the PFS model K=1. The mean SNR improvement measured for the TVAR model was 6.27 dB and the SNR improvement measured for the PFS model was 7.21 dB. The PFS model provided slightly higher but consistent SNR improvements for this setup. Halving the value of the random walk parameters δ_f and δ_b produced a mean SNR improvement of 7.12 dB and doubling them produced 7.16 dB. The value of these parameters could be changed at least an order of magnitude and still produce higher SNR improvements than that of the TVAR model which favors the PFS model as being a better speech model than the TVAR model.

In the particle filter the unknown parameters are augmented to the state vector. In this way the particle filter provides estimates of the unknown parameters. Using *Praat* [6] the formant frequency tracks were extracted from the clean speech. The 'true' $F1$ formant frequency is illustrated in the upper plot of Fig. 3 together with the estimated $F1$ formant frequency tracks using the TVAR model and the PFS model. The estimated formant frequency tracks were obtained by averaging the estimates from 7 repeated experiments. By using the inverse mapping in eqs. (6) and (7) the formant frequency track for the TVAR model was computed from the estimated AR coefficients. As is evident from Fig. 3a the PFS model provides a much better estimate of the formant frequency than the TVAR model. It is also seen from Fig. 3b that the PFS model provides a more smooth and accurate estimate of the AR coefficients. Next, performance was measured for the sound with $F1$ and $F2$ using the same conditions as for the sound with $F1$ only. The PFS model and the TVAR model gave 6.24 dB and 5.41 dB mean SNR improvements, respectively. The PFS model provided slightly higher and consistent SNR improvements for this setup also. The estimated formant frequency tracks using the models are seen in Fig. 4. This experiment illustrated even more the convenience of the PFS model over the TVAR model in that it's much more straight forward to use the properties of

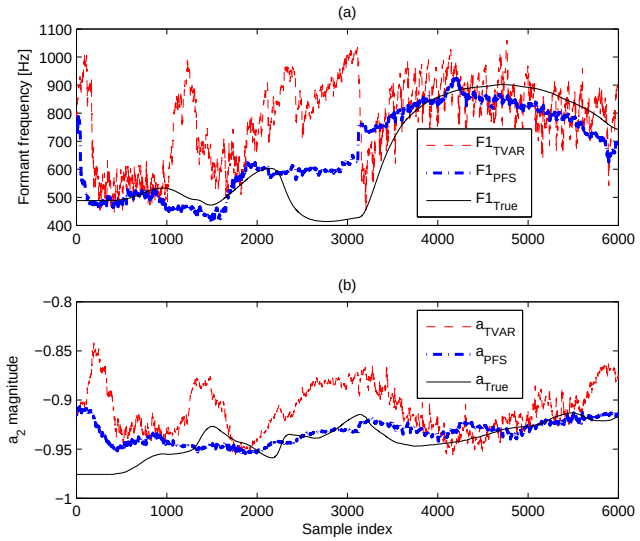

Fig. 3. (a) Estimated $F1$ formant frequency track using the TVAR and PFS models and the 'true' formant frequency extracted from clean speech. (b) Estimated tracks of the a_2 AR coefficient together with the 'true' values.

the PFS model to ensure reasonable behavior of particle paths. It is for instance more cumbersome to initialize the TVAR model so that the formant frequencies of the particle paths get in range with the formants of the sound. It is also significantly more cumbersome to ensure that the particle paths of the TVAR model remain within the limits of the range of $F1$ and $F2$. If this is not ensured the estimated spectrum of the sound using the TVAR model can have a low-pass characteristic or a single peak instead of two peaks. Performance was then

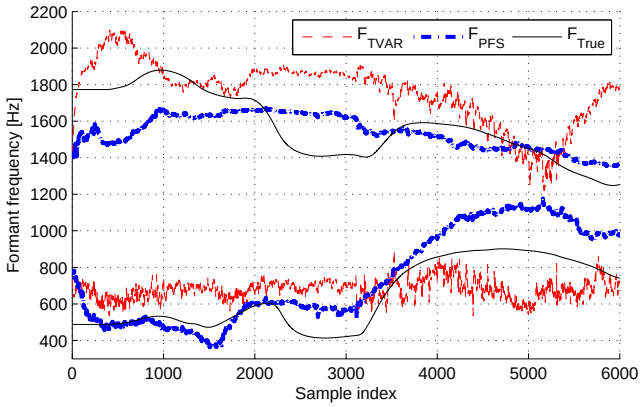

Fig. 4. Estimated $F1$ and $F2$ formant frequency tracks using the TVAR and PFS models and the 'true' formant frequencies extracted from clean speech.

measured for the two fullband waveforms shown in Fig. 2. SNR improvements were measured for 4 different SNRs using both models and same conditions as in previous experiments. The results are seen in Table I. The PFS model produced higher SNR improvements for both sounds and all 4 SNRs. It's also seen that there is a negative correlation between measured dB SNR improvement and SNR. As a last experiment particle filtering was performed on the full length TIMIT waveforms in order to test the quality of the enhanced speech signals. Only one run on each sound was made and the variance of

TABLE I

MEASURED DB SNR IMPROVEMENTS FOR THE TVAR AND PFS MODELS FOR 4 DIFFERENT SNRS USING THE SI1466 AND SX111 TIMIT SOUNDS.

	si1466				sx111			
Model	0	5	10	20	0	5	10	20
TVAR	5.20	3.06	1.03	0.24	3.80	1.59	0.92	−0.02
PFS	5.69	4.54	3.07	1.43	4.82	2.92	1.62	0.42

the noise was made time-varying and the particle filtering changed accordingly. The listening tests revealed that the quality of the enhanced speech signals was rather poor for both models and notable artifacts were introduced. It is believed that an important factor contributing to the relative poor speech quality for both models is their shortcoming in accurately modelling the excitation source for voiced speech.

VI. CONCLUSION & OUTLOOK

We have proposed a new parametrization of a time-varying auto-regressive speech model and used particle filtering for inference in a noise reduction set-up. The performance of a proposed speech model was compared to that of a speech model parameterized by auto-regressive coefficients for the application of speech enhancement [1]. The results from a number of experiments showed that the proposed model provided higher SNR estimates of the speech over a large interval of the random walk parameters of the particle filter and more accurate and smooth estimates of the model parameters were obtained as well which favors the proposed model as a better model for speech. Listening tests reveal that despite of higher SNR improvements there is still room for substantial improvement. Work is currently being made on both improving the proposed model and the inference. Voiced speech is not well modelled by a random excitation signal. We expect that a harmonic excitation term would be beneficial. It is possible to improve the inference by drawing samples from a proposal distribution extending some finite time in the past. Setting this time-horizon to be equal to a few time the auto-correlation of the process will allow us to achieve smoothing (i.e. full posterior rather than filtering) estimates with very little time delay (Ferkinghoff-Borg, Lehn-Schiøler and Winther, personal communication).

ACKNOWLEDGMENT

The authors would like to thank Jaco Vermaak for making the code used in [1] available to them.

REFERENCES

[1] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 173–185, 2002.

[2] A. Doucet, S. Godsill, and M. West, "Monte carlo filtering and smoothing with application to time-varying spectral estimation," in *In Proc. ICASSP*, Istanbul, Turkey, June 2000, pp. 701–704.

[3] K. Hermansen and U. Hartmann, "Fbg model based low rate coding of speech," in *Proceedings of NORSIG 2002 : The 5th IEEE Nordic Signal Processing Symposium*, Norway, Oct. 2002.

[4] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York, USA: Springer-Verlag, New York, 1976.

[5] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, pp. 971–995, 1980.

[6] P. Boersma and D. Weening. (2006) Praat: doing phonetics by computer (version 4.4.24). Compter program. [Online]. Available: http://www.praat.org/